White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks

R. Guerrero[a,*,1], C. Qin[a,1], O. Oktay[a], C. Bowles[a], L. Chen[a], R. Joules[b], R. Wolz[b,a], M.C. Valdés-Hernández[c], D.A. Dickie[c], J. Wardlaw[c], D. Rueckert[a]

[a] Department of Computing, Imperial College London, UK
[b] IXICO plc., UK
[c] UK Dementia Research Institute at The University of Edinburgh, Edinburgh Medical School, 47 Little France Crescent, Edinburgh EH16 4TJ, UK

ARTICLE INFO

ABSTRACT

White matter hyperintensities (WMH) are a feature of sporadic small vessel disease also frequently observed in magnetic resonance images (MRI) of healthy elderly subjects. The accurate assessment of WMH burden is of crucial importance for epidemiological studies to determine association between WMHs, cognitive and clinical data; their causes, and the effects of new treatments in randomized trials. The manual delineation of WMHs is a very tedious, costly and time consuming process, that needs to be carried out by an expert annotator (e.g. a trained image analyst or radiologist). The problem of WMH delineation is further complicated by the fact that other pathological features (i.e. stroke lesions) often also appear as hyperintense regions. Recently, several automated methods aiming to tackle the challenges of WMH segmentation have been proposed. Most of these methods have been specifically developed to segment WMH in MRI but cannot differentiate between WMHs and strokes. Other methods, capable of distinguishing between different pathologies in brain MRI, are not designed with simultaneous WMH and stroke segmentation in mind. Therefore, a task specific, reliable, fully automated method that can segment and differentiate between these two pathological manifestations on MRI has not yet been fully identified. In this work we propose to use a convolutional neural network (CNN) that is able to segment hyperintensities and differentiate between WMHs and stroke lesions. Specifically, we aim to distinguish between WMH pathologies from those caused by stroke lesions due to either cortical, large or small subcortical infarcts. The proposed fully convolutional CNN architecture, called uResNet, that comprised an analysis path, that gradually learns low and high level features, followed by a synthesis path, that gradually combines and up-samples the low and high level features into a class likelihood semantic segmentation. Quantitatively, the proposed CNN architecture is shown to outperform other well established and state-of-the-art algorithms in terms of overlap with manual expert annotations. Clinically, the extracted WMH volumes were found to correlate better with the Fazekas visual rating score than competing methods or the expert-annotated volumes. Additionally, a comparison of the associations found between clinical risk-factors and the WMH volumes generated by the proposed method, was found to be in line with the associations found with the expert-annotated volumes.

1. Introduction

1.1. Clinical motivation

White matter hyperintensities (WMH), referred to in the clinical literature as leukoaraiosis, white matter lesions or white matter disease (Wardlaw et al., 2013), are a characteristic of small vessel disease (Wardlaw and Pantoni, 2014) commonly observed in elderly subjects on fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) images, which, as the name suggests, they appear as hyperintense regions. Moreover, stroke lesions of cortical, large subcortical (striato-capsular) or small subcortical infarct origin can also often appear as hyperintense regions in FLAIR MR images and can coexist and coalesce with WMHs. The accurate assessment of WMH burden is of crucial importance for epidemiological studies to determine associations between WMHs, cognitive and clinical data. Similarly, it would help

discover their causes, and the effects of new treatments in randomized trials. In the assessment of WMH burden it is important to exclude stroke lesions as they have different underlying pathologies, and failure to account for this may have important implications for the design and sample size calculations of observational studies and randomized trials using WMH quantitative measures, WMH progression or brain atrophy as outcome measures (Wang et al., 2012). One of the most widely used metrics to assess WMH burden and severity is the Fazekas visual rating scale (i.e. score) (Fazekas et al., 1987). In this scale, a radiologist visually rates deep white matter and peri-ventricular areas of a MR scan into four possible categories each depending on the size, location and confluence of lesions. The combination of both deep white matter and peri-ventricular ratings yields a combined zero to six scale. In the vast majority of clinical trials and in general clinical practice visual rating scores are used (such as the Fazekas score). WMHs are very variable in size, appearance and location, and therefore the categorical nature of the Fazekas scale has limitations for studying their progression in relation with other clinical parameters. WMH volume has been demonstrated to correlate with severity of symptoms, progression of disability and clinical outcome (Bendfeldt et al., 2010; Chard et al., 2002; Löuvbld et al., 1997). Accordingly, determining WMH volume has been of interest in clinical research as well as in clinical trials on disease-modifying drugs (Löuvbld et al., 1997; van Gijn, 1998; Brott et al., 1989; SPIRIT, 1997). For some studies, lesions have been traced manually (sometimes with the help of semi-automated tools for contour detection) slice by slice. This process can easily become prohibitively expensive for even moderately large datasets. It is therefore obvious that the accurate automatic quantification of WMH volume would be highly desirable, as this will undoubtedly lead to savings in both time and cost. Recently, several automated and semi-automated methods have been put forward to address the coarseness of the visual assessments (e.g. Fazekas score), as well as the dependence on highly qualified experts to perform such assessments. These methods can be broadly classified into *supervised*, when a training "gold-standard" is available (Van Nguyen et al., 2015; Ghafoorian et al., 2016), i.e. when one or more human experts have annotated data, *unsupervised*, when no such gold-standard exists (Ye et al., 2013; Cardoso et al., 2015; Bowles et al., 2016), and *semi-supervised*, when only a small portion of available data has been expertly annotated (Kawata et al., 2010; Qin et al., 2016). However, despite the number of proposed methods, no automated solution is currently widely used in clinical practice and only a few of them are publicly available (Shiee et al., 2010a; Damangir et al., 2012; Schmidt et al., 2012). This is partly because lesion load, as defined in most previously proposed automatic WMH segmentation algorithms, does not take into account the contribution of strokes lesion, as these methods are generally unable to differentiate between these two types of lesions.

1.2. Related work

In the following we review existing methods and challenges that are related to our work, especially on Multiple sclerosis (MS), WMH and stroke lesion segmentation in MR imaging. Additionally, some more general CNN segmentation approaches that share architectural similarities with the method we propose here are also reviewed in this section. Over the last few years, there has been an increased amount of research going on in these areas (García-Lorenzo et al., 2013; Caligiuri et al., 2015; Maier et al., 2017; Rekik et al., 2012). Although some of the methods mentioned here were proposed for segmenting different pathologies rather than the ones we explore in this work, they can in fact be applied to different tasks. As mentioned before, these methods can be broadly classified into *unsupervised*, *semi-automatic*, *semi-supervised* and *supervised*, depending on the amount of expertly annotated data available.

1.2.1. Unsupervised segmentation

Unsupervised segmentation methods do not require labeled data to perform the segmentation. Most of these approaches employ clustering methods based on intensity information or some anatomical knowledge to group similar voxels into clusters, such as fuzzy C-means methods (Gibson et al., 2010), EM-based algorithms (Dugas-Phocion et al., 2004; Forbes et al., 2010; Kikinis et al., 1999) and Gaussian mixture models (Freifeld et al., 2009; Khayati et al., 2008). Some of the probabilistic generative models of the lesion formation for stroke lesion segmentation were also designed, such as Forbes et al. (2010); Derntl et al. (2015). Forbes et al. (2010) proposed a Bayesian multi-sequence Markov model for fusing multiple MR sequences to robustly and accurately segment brain lesions. Derntl et al. (2015) proposed to combine standard atlas-based segmentation with a stroke lesion occurrence atlas, in a patient-specific iterative procedure. Some authors have also proposed to model lesions as outliers to normal tissues. Van Leemput et al. (2001) employed a weighted EM framework in which voxels far from the model were weighted less in the estimation and considered potential lesions. Weiss et al. (2013) proposed to use dictionary learning to learn a sparse representation from pathology free brain T1-weighted MR scans and then applied this dictionary to sparsely reconstruct brain MR images that contain pathologies, where the lesions were identified using the reconstruction error. Additionally, several works have also focused on exploiting the fact that WMHs are best observed in FLAIR MR images, while being difficult to identify in T1-weighted MR images. Some of these methods rely on generating a synthetic FLAIR image based on observed T1-weighted MR image using random forests (Ye et al., 2013), generative mixture-models (Cardoso et al., 2015), support vector regression (SVR) (Bowles et al., 2016) or convolutional neural networks (CNN) (Van Nguyen et al., 2015). Both synthetic (healthy looking) and real FLAIR (with pathologies) images are then compared to detect any abnormalities. Other method like lesion-TOADS (Shiee et al., 2010b) combines atlas segmentation with statistical intensity modeling to simultaneously segment major brain structures as well as lesions. The lesion growth algorithm (LGA), proposed by Schmidt et al. (2012) and part of SPM's LST toolbox (www.statistical-modelling.de/lst.html), constructs a conservative lesion belief map with a pre-chosen threshold (κ), followed by the initial map being grown along voxels that appear hyperintense in the FLAIR image. In essence, LGA is a self-seeded algorithm and it tends to have difficulties detecting subtle WMHs. An important drawback of all these methods is that they are in fact abnormality detection algorithms and not specifically WMH segmentation methods, hence in principle they detect any pathology, whether or not is a WMH-related pathology.

1.2.2. Semi-automatic and semi-supervised segmentation

Several semi-automatic algorithms proposed in the literature for WMH segmentation rely on region growing techniques that require initial seed points to be placed by an operator. Kawata et al. (2010) introduced a region growing method for adaptive selection of segmentation by using a SVM with image features extracted from initially identified WMH candidates. Itti et al. (2001) proposed another region growing algorithm that extracts WMHs by propagating seed points into neighboring voxels whose intensity is above an optimized threshold. The process iterates until convergence, i.e. all voxels above the threshold that are connected to the initial seed point had been annotated. Aside from the drawback of requiring per image expert inputs, semi-automatic methods have the additional potential drawback that seeds points could easily be selected in obvious regions, while the biggest challenge of WMH segmentation can arguably be found in the more confusing border regions. Qin et al. (2016) proposed a semi-supervised algorithm that optimizes a kernel based max-margin objective function which aims to maximize the margin averaged over inliers and outliers while exploiting a limited amount of available labeled data. Although theoretically interesting and well motivated, the problem of transferring useful knowledge from unlabeled data to a task defined by

partially annotated data remains a challenge and an open field of research in its own right. Hence, in practice, semi-supervised WMH segmentation methods, even though they still require some expert input, tend to underperform when compared to supervised methods, even when the later are trained with only a modest amount of data.

1.2.3. Supervised segmentation

Supervised methods for lesion segmentation have also been well researched. Classical supervised machine learning methods such as k-nearest neighbors (kNN) (Anbeek et al., 2004a), Bayesian models (Maillard et al., 2008), support vector machines (SVM) (Lao et al., 2008), and random forests (Geremia et al., 2011) have been well studied in MS segmentation. For stroke lesion segmentation, pattern classification techniques to learn a segmentation function were also employed in Prakash et al. (2006); Maier et al. (2014); Maier et al. (2015b); Maier et al. (2015a). The lesion prediction algorithm (LPA) (Schmidt, 2017), implemented in SPM's LST toolbox, has been shown to produce consistently good performance and in many cases is considered a robust gold standard for this problem. LPA is described as a logistic regression model, where binary lesion maps of 53 MS patients were used as response values. Additionally, as covariates to this model a lesion belief map similar to those from LGA (Schmidt et al., 2012) was used in combination with a spatial covariate that takes into account voxel specific changes in lesion probability. Recently, Ithapu et al. (2014) proposed using SVMs and random forests in combination with texture features engineered by texton filter banks for WMH segmentation task. Brain intensity abnormality classification algorithm (BIANCA) (Griffanti et al., 2016), a fully automated supervised method based on kNN algorithm, was also proposed for WMH segmentation. An interesting work proposed by Dalca et al. (2014) used a generative probabilistic model for the differential segmentation of leukoaraiosis and stroke by learning the spatial distribution and intensity profile of each pathology, which shares the same application purpose with the work proposed here.

More recently, CNNs have been put forward to replace the inference step in many computer vision related tasks (Girshick et al., 2014; Long et al., 2015; He et al., 2016a; Dong et al., 2016), with current state-of-the-art methods in many fields being dominated by CNN frameworks. CNNs have been shown to have enough capacity to model complex nonlinear functions capable of performing multi-class classification tasks such as those required for the description and understanding of highly heterogeneous problems, such as brain lesion segmentation (Brosch et al., 2015; Birenbaum and Greenspan, 2016; Kamnitsas et al., 2015; Kamnitsas et al., 2017; Valverde et al., 2016; McKinley et al., 2016). For instance, Brosch et al. (2015) proposed a deep convolutional encoder network which combines feature extraction and segmentation prediction on MS lesions. Their work was later extended to a 3D deep encoder network with shortcut connections, which consistently outperformed other methods across a wide range of lesion sizes (Brosch et al., 2016). Kamnitsas et al. (2017) proposed a network architecture with two parallel convolutional pathways that processes the 3D patches at different scales followed by a 3D densely connected conditional random field (CRF). Although the method was originally proposed for ischemic stroke, tumor and brain injury segmentation on MR images, it can be easily adapted for different tasks using their provided package DeepMedic[2]. Similarly, Ghafoorian et al. (2016) proposed a CNN architecture that considered multi-scale patches and explicit location features while training, and later was extended to consider non-uniform patch sampling (Ghafoorian et al., 2016). Their best performing architecture shares a similar design with the architecture proposed by Kamnitsas et al. (2015, 2017), in which it trained independent paths of convolutional layers for each scale.

Using multi-resolution inputs (Kamnitsas et al., 2017; Ghafoorian

---

[2] https://github.com/Kamnitsask/deepmedic

et al., 2016; Ghafoorian et al., 2016) can increase the field of view with smaller feature maps, while also allowing more non-linearities (more layers) to be used at higher resolution, both of which are desired properties. However, down-sampling patches has the drawback that valuable information is being discarded before any processing is done, and since filters learned by the first few layers of CNNs tend to be basic feature detectors, e.g. lines or curves, different paths risk capturing redundant information. Furthermore, although convolutions performed in 3D as in Kamnitsas et al. (2017) intuitively make sense for 3D volumetric images, FLAIR image acquisitions are actually often acquired as 2D images with large slice thickness and then stacked into a 3D volume. Further to this, gold standard annotations, such as those generated by trained radiologists (e.g. WMH delineation or Fazekas scores) are usually derived by assessing images slice by slice. Thus, as pointed out by Ghafoorian et al. (2016), 3D convolutions for FLAIR MR image segmentation are in fact less intuitive.

Some other works on CNN segmentation which are relevant to our work, though not on brain lesion segmentation, include Long et al. (2015) and Ronneberger et al. (2015). Long et al. (2015) proposed to segment natural images using a fully convolutional network that supplemented the output of a gradually contracting network with features from several of its levels of contraction through up-sampling. Similar to Long et al. (2015); Ronneberger et al. (2015) used a U-shaped architecture (U-net) to segment microscopical cell images. The architecture symmetrically combined a contracting and expanding path via feature concatenations, in which up-sampling operations were realized with trainable kernels (deconvolution or transposed convolution). Both of these networks form the foundation of the architecture later proposed in this work.

1.2.4. Challenges

There are several challenges being held on brain lesion segmentation in recent years. For instance, the MS lesion segmentation challenge 2008 (http://www.ia.unc.edu/MSseg/) had the goal of the direct comparison of different 3D MS lesion segmentation techniques. Data used in this challenge consisted of 54 brain MR images from a wide range of patients and pathology severity. The 2015 Longitudinal MS Lesion Segmentation Challenge (http://iacl.ece.jhu.edu/index.php/MSChallenge) aimed to apply automatic lesion segmentation algorithms to MR neuroimaging data acquired at multiple time points from MS patients. The ischemic stroke lesion segmentation (ISLES) challenge (http://www.isles-_challenge.org/) has been held since 2015, which aims to provide a platform for a fair and direct comparison of methods for ischemic stroke lesion segmentation from multi-spectral MRI image, and asked for methods that allow the prediction of lesion outcome based on acute MRI data. More recently, a WMH segmentation challenge (http://wmh.isi.uu.nl/) was held aiming to directly compare methods for the automatic segmentation of WMH of presumed vascular origin, with data used in this challenge acquired from five different scanners from three different vendors in three different hospitals.

1.3. Contributions

In this work we aim to address some of the shortcomings mentioned before and propose to use a CNN to segment and differentiate between WMH-related pathology and strokes. Specifically, we task ourselves with distinguishing between WMH pathologies from those pathologies originating due to stroke lesions that result from either cortical or subcortical infarcts. For this, a CNN with an architecture inspired by U-net (Ronneberger et al., 2015), originally used to segment neuronal structures in electron microscopic stacks, is proposed. The architecture consists of an analysis path that aims to capture context and a symmetric synthesis path that gradually combines analysis and synthesis features to ultimately enable precise localization. The proposed CNN architecture is trained with large high-resolution image patches and is able to extract high- and low-level features through a single path, thus

avoiding filter learning redundancy. Different to Ronneberger et al. (2015), in the work proposed here we replace convolutions with residual elements (He et al., 2016a) and concatenations used in skip connections in the U-net architecture with summations to reduce model complexity. Residual architectures have been shown to ease gradient back-propagation flow, and hence improve optimization convergence speed and allow for deeper network training. An important contribution of this work deals with data sampling for training. Due to the large class imbalance present in WMH segmentation, data sampling for training requires careful consideration, an issue that has received recent research focus due to its influence on the precision of segmentation (Kamnitsas et al., 2017). Here, to mitigate class imbalance, training is done using patches, rather than dense training on whole images. Further to this, we sample patches that always contain WMH and randomly shift the central location so that WMH can occur anywhere in the patch and not necessarily include the center. As argued before, the proposed CNN architecture is designed for 2D images and it is trained with 2D image patches. Furthermore, we experiment with multi-channel inputs to evaluate the added benefit of adding T1 MR scans and white matter and/or cerebro-spinal track probability maps. The proposed architecture, which we refer as uResNet, can be visualized in Fig. 1.

2. Methods

CNNs represent a versatile class of machine learning models that can be trained to predict voxel-wise semantic labels on images. This is achieved by learning a mapping function $f(\Theta, x) \rightarrow y$, parametrized by Θ, that transforms voxel level image intensity x to a desired label space or image segmentation $y \in Y$. Such mapping function $f(\Theta, x)$ is modeled by a series of L convolution and non-linearity operations, with each element in this sequence generally referred to as a layer. Each layer l produces a set of features maps H_l. Here, the convolutional kernel of layer l that produces the jth feature map is parametrized as $w_l^{j,k}$, where k refers to the kth feature map of H_{l-1}. The solution to this problem estimates a conditional distribution $p(y|x)$ that minimizes the loss function Ψ (see Section 2.2) defined by y and its estimate $f(\Theta, x)$. After each layer l a set of feature maps or intermediate representations h_j^l is obtained. In this work, non-linearities are defined as rectified linear units (ReLU) (Nair and Hinton, 2010). Intermediate feature maps are computed as convolutions between the convolution kernels $w_l^{j,k}$ and the layers' input as

$$h_l^j = \max\left(0, \sum_{k=1}^{J_{l-1}} h_{(l-1)}^k * w_l^{j,k}\right). \tag{1}$$

Here $*$ denotes the convolution operator, $h_0^j = x$, and J_{l-1} is the number feature maps in layer $l-1$, with J_0 being the number of input

channels.

In addition to the sequence of convolution and non-linearity operations mentioned, in the work presented here, residual units or residual elements (ResEle) (He et al., 2016a) are employed to reformulate the previous mapping function as $f(\theta_l, H_{l-1}) + W_l H_{l-1} \rightarrow H_l$, where W_l performs a linear projection that matches the number of feature maps in layer $l-1$ to those in layer l. Fig. 1 bottom-right shows the form of ResEle used in this work. Furthermore, to decrease the number of parameters (and control over-fitting) associated with an increase network field-of-view max-pooling layers are employed. Max-pooling operates independently on each input feature map where all but the maximum valued activation within a support region is discarded, the same is repeated at every strided location. Support region and stride in this work were set to 2×2 and 2, respectively, effectively down-sampling by a factor of two after every max-pool layer.

2.1. Network architecture

Defining a CNN's architecture requires that careful consideration of the task is set out to achieve. Important aspects that must be taken into account are the network's field of view or receptive field and its capacity or complexity. In the architecture proposed here we follow the suggestions of Simonyan and Zisserman (2015) and use only small (3×3) kernels. This allows an increased non-linearity capacity with a lower number of parameters needed for the same receptive field.

The architecture proposed here follows a U-shaped architecture. Furthermore, no fully connected layers are used, thus it is a fully convolutional network, and hence even though it is trained with image patches, inference can be performed on whole images in one single feed forward pass without any need of architectural changes. In total our architecture is composed of 12 layers with ~1 M trainable parameters: 8 residual elements, 3 deconvolution layers, and one final convolution layer that converts the feature maps to the label space. Here, the last layer's feature maps H_L are passed to an element-wise softmax function that produces pseudo class probability maps as
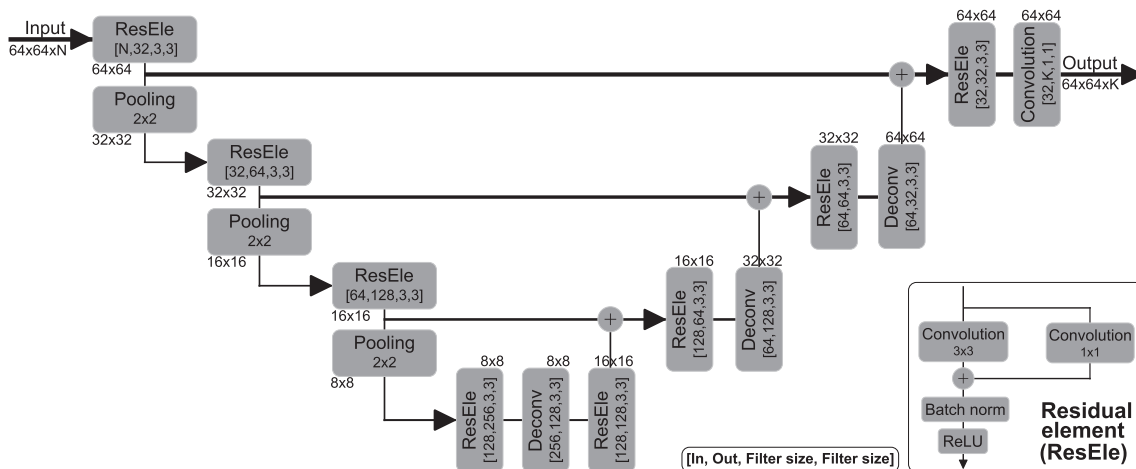
$$\rho_c(H_L) = \frac{\exp(H_L)}{\sum_{c=1}^{C} \exp(H_L)} \quad \forall \, c, \tag{2}$$

where c denotes class and C the total number of classes.

This in essence yields a class-likelihood for each voxel in the image, and its output, in combination with a loss function (described in Section 2.2), is optimized through the back-propagation algorithm.

2.2. Loss function and class imbalance

In general terms, a loss function maps the values of one or more

Fig. 1. Proposed u-shaped residual network (uResNet) architecture for WMH segmentation and differentiation.

variables onto a real number that represents some "cost" associated with an event. Loss functions defined for classification tasks are functions that calculate a penalty incurred for every incorrect prediction. As mentioned before, casting a semantic segmentation task as a voxel-wise classification problem tends to lead to significant class imbalances. Loss functions can be defined in such a way that they take class imbalance into account. Here, we will detail a classical loss function that does not take into account class imbalance as well as several recently proposed loss functions that either directly or indirectly take into account class imbalance, as they will be subject of investigation.

In the context of the work presented here, let us define a training set of samples $X = \{x_1,\ldots,x_P\}$, where each $x_p = \{x_{(p,1)},\ldots,x_{(p,V)}\}$ are image patches extracted from in-plane FLAIR (and/or additional modalities) axial slices that will be treated as independent samples during training. Here, P is the total number of patches available and V the total number of voxels per patch. Additionally, let us also define voxel level labels as one-hot encoded variables $y_{p,v}$ associated with each voxel $x_{p,v} \in X$. Let us consider $Y \in \mathbb{N}^C$ a one-hot encoded label space, where the class of each voxel in $x_{p,v}$ is given by a C-length vector $y_{p,v}$ of all zeros except for one at position c which indicates the associated label. However, let us simplify notation for the following loss equations by re-indexing all voxels in X and their corresponding label as x_n and y_n, respectively. Here, $n = \{1,\ldots,N\}$ and $N = P * V$ is the total number of voxels from all patches in X. Therefore, the problem of estimating the mapping function $f(\Theta,x_n)$ can be defined as the minimization of a loss function that works with the pseudo probabilities obtained from Eq. (2).

A popular loss function for classification tasks, such as the one tackled here, is the categorical cross-entropy which aims to maximize the log likelihood of the data or, equally, minimize the cross-entropy via the following loss function

$$\Psi = -\sum_{n=1}^{N} y_n \log(f(\Theta, x_n)). \tag{3}$$

Classical cross-entropy does not take into account class imbalances in the data which might lead to learning biased predictors. A simple approach to deal with class imbalance that has been proposed for CNN segmentation, is to modify the aggregation of categorical cross-entropy given in Eq. (3), by weighting voxels that belong to different classes differently. This modification aims to give more weight to under-represented classes, while weighting down over-represented ones, and can be written as

$$\Psi = -\sum_{n=1}^{N} y_n \log(f(\Theta, x_n))\omega(y_n). \tag{4}$$

where $\omega(y_n)$ is the weight associated to class of y_n.

Wu et al. (n.d.) recently proposed a simple modification of the categorical cross-entropy by dropping, or ignoring, the loss contribution of elements whose correct class prediction was above a certain threshold τ. This has the effect of placing more emphasis on previous mistakes, thus focusing the learning process on "harder" (and arguably more valuable) examples during training. Dubbed online bootstrapped categorical cross-entropy, this loss function can be written as

$$\Psi = -\sum_{n=1}^{N} y_n \log(\varphi_n) \tag{5}$$

where $\varphi_n = \{1 \text{ if } f(\Theta,x_n) > \tau, f(\Theta,x_n) \text{ otherwise}\}$.

The Dice coefficient is defined on a binary space and aims at maximizing the overlap between regions of the same class. This makes it a popular and natural choice of metric when comparing binary segmentation labels. However, it is non-differentiable, making its optimization with the back-propagation algorithm not possible. Recently, the winning team of the Second Annual Data Science Bowl[3], proposed using a

pseudo Dice coefficient as loss function, that can be written as

$$\Psi = 1 - \frac{1}{C} \sum_{c=1}^{C} \left(\frac{2 \sum_{n=1}^{N} (y_n^c f(\Theta, x_n)^c)}{\sum_{n=1}^{N} f(\Theta, x_n)^c + \sum_{n=1}^{N} y_n^c} \right). \tag{6}$$

Here, the predicted binary labels are replaced by continuous softmax outputs and averaged across all labels C, and where $f(\Theta,x_n)^c$ denotes the softmax prediction of class c. Aggregating Dice coefficients from C different classes as an average, has the additional effect of normalizing the per-class loss contribution.

2.3. Data sampling and class imbalance

Generally, in the segmentation of pathologies, healthy tissue is present in far larger quantities than pathological. For example, in WMH segmentation the number of voxels labeled as WMH (regardless of the underlying pathology) is very small compared to those labeled background/healthy tissue, which leads to a significant class imbalance ($\sim 99.8\%$ of the voxels in the dataset used in this work are labeled as background/healthy tissue in our training set). Hence, although dense training (where whole images or slices are used) is a staple in computer vision with natural images (Long et al., 2015), it is less intuitive for WMH segmentation. Therefore, patch sampling is used in this work in order to alleviate the class imbalance problem. There are several techniques that could be used to sample patches for training. For example half of the samples could be extracted from locations centered on healthy tissue and half centered on WMH tissue (Kamnitsas et al., 2017), however this strategy does little for the class imbalance when large patches are being considered, as individual patches tend to still be highly class imbalanced at a voxel level. Another option, is to sample patches centered at WMH locations only, which in fact reduces the healthy tissue class to $\sim 90\%$. However, both strategies, in combination with the proposed architecture that has a field of view comparable to sample size, would lead to a location bias, where WMHs are always expected in the center of the patch. Instead, we propose that after defining a random subset of WMH voxels from which to extract training patches, a random shift $\Delta_{x,y}$ of up to half the patch size be applied in the axial plane before patch extraction to augment the dataset. Fig. 2 details this procedure. It is important to point out that the location sensitivity mentioned here, is generally not an issue with dense training in natural images, where different classes can either appear anywhere in a scene (e.g. a face might be located anywhere), or class location gives a meaningful description (e.g. sky tends to be in the upper part of a scene). This problem only occurs when sampling patches from training images in a systematic way, such as proposed here.

3. Data

The proposed methodology was evaluated using a subset of 167 images from 250 consecutive patients who presented themselves to a hospital stroke service with their first clinically evident non-disabling lacunar or mild cortical ischemic stroke (Valdés Hernández et al., 2015a). Diabetes, hypertension, and other vascular risk factors were not criteria for exclusion. However, patients with unstable hypertension or diabetes, other neurological disorders, major medical conditions including renal failure, contraindications to MRI, unable to give consent, those who had hemorrhagic stroke, or those whose symptoms were resolved within 24 h(i.e., transient ischemic attack) were excluded. The subset of 167 subjects considered in this work consisted of those for which all WMH and stroke lesions were delineated (see Section 3.1.1) as different annotation classes, i.e. those that contained strokes but were not labeled as such were excluded. In this work, stroke lesions included both old and recent lesions as defined in Valdés Hernández et al. (2015a), which in turn are either of cortical or sub-cortical nature.

A subset of 126 from the 167 subjects used, contained additional complete clinical and demographic data. Information included risk
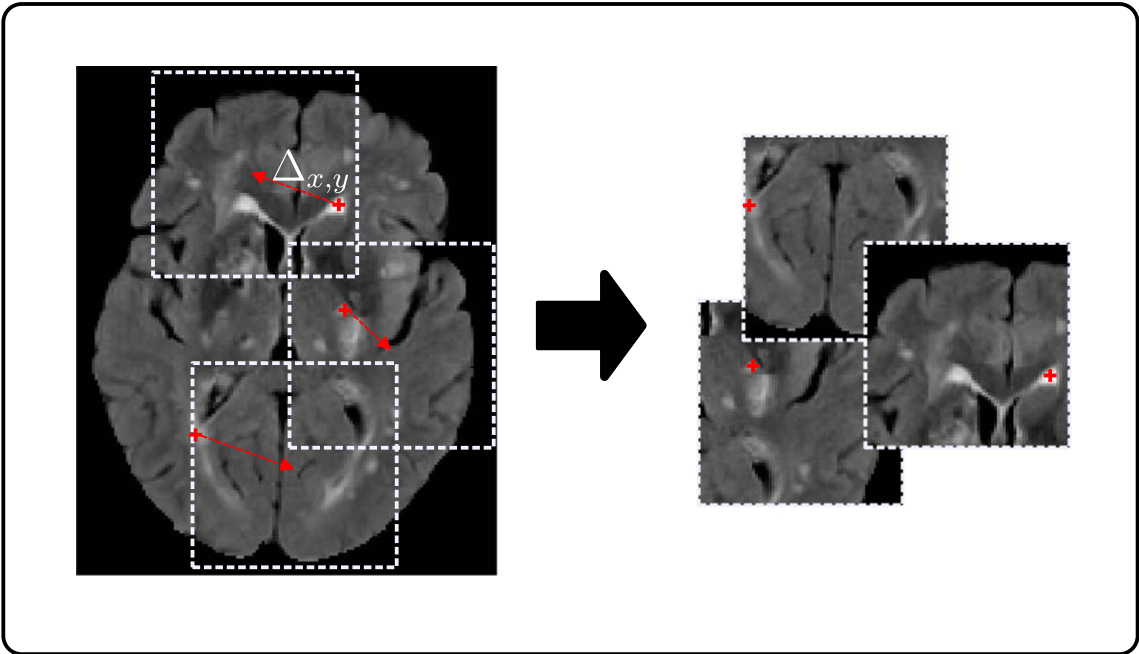
---

[3] https://www.kaggle.com/c/second-annual-data-science-bowl

Fig. 2. Training patch sampling.

factors and clinical assessments such as: age, sex, reported diabetes, reported hypertension, reported hyperlipidaemia, reported smoking, mini mental state examination (MMSE), systolic blood pressure (SBP), diastolic blood pressure (DBP), total cholesterol, peri-ventricular Fazekas score (PV-Fazekas), deep white matter Fazekas score (D-Fazekas), deep atrophy volume (deepAtrophyVol), basal ganglia enlarged peri-vascular spaces (BGPVS) score, centrum semiovale enlarged peri-vascular spaces (CSPVS) score, old stroke lesion (oldLes) present, and total number of micro-bleeds (micrBld).

3.1. MRI acquisition

All image data was acquired at the Brain Research Imaging Centre of Edinburgh (http://www.bric.ed.ac.uk) on a GE Signa Horizon HDx 1.5 T clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer-supplied eight-channel phased-array head coil. Details of the protocols used for acquiring the data are given in Table 1, and their rationale is explained in Valdés Hernández et al. (2015a). Although several imaging sequences were acquired, only T1 and FLAIR MR images were used for this study. Of the 167 subjects considered in this work 35 were acquired under protocol

Table 1
MR imaging sequence details for the three acquisition protocols used.

Protocols	Protocol 1	Protocol 2	Protocol 3
TR/TE/TI (ms) T1	9/440	9.7/3.984/500	
TR/TE/TI (ms) FLAIR	9002/147/2200	9000/140/2200	
Pixel bandwidth (KHz)	125 (T1) 122.07(FLAIR)	15.63 (T1) 15.63 (FLAIR)	
Matrix	256 × 192	256 × 216 (T1) 384 × 224 (FLAIR)	192 × 192 (T1) 256*256(FLAIR)
No. slices	20	256 (T1) 28 (FLAIR)	160 (T1) 40 (FLAIR)
Slice thickness (mm)	5	1.02 (T1) 5 (FLAIR)	1.3 (T1) 4 (FLAIR)
Inter-slice gap (mm)	1.5	1	0
Voxel size (mm³)	0.94 × 0.94x6.5	1.02 × 0.9x1.02 (T1) 0.47 × 0.47x6 (FLAIR)	1.3 × 1.3x1(T1) 1 × 1x4 (FLAIR)

1, 83 under protocol 2 and 49 under protocol 3.

3.1.1. Image pre-processing and gold standard annotations

All image sequences (from each patient) were co-registered using FSL-FLIRT (Jenkinson et al., 2002) and mapped to the patient's FLAIR space. WMH from MR images that were acquired under protocol 2 (Table 1) were delineated using Multispectral Coloring Modulation and Variance Identification (MCMxxxVI). Described in Valdés Hernández et al. (2015a,b), MCMxxxVI is based on the principle of modulating, or mapping, in red/green/blue color space two or more different MRI sequences that display the tissues/lesions of the brain in different intensity levels, before employing Minimum Variance Quantization (MVQ) as the clustering technique to segment different tissue types. Here, MCMxxxVI considers WMH those hyperintense signals that simultaneously appear in all T2-weighted based sequences. WMH from MR images acquired under the protocols 1 and 3 were delineated via a human corrected histogram-based threshold of the FLAIR sequence. Stroke lesions (old and recent) were separately extracted semi-automatically by thresholding and interactive region-growing, while guided by radiological knowledge, on FLAIR (if ischemic) or T2*W (if hemorrhagic) (Valdés Hernández et al., 2015a,b). Their identification procedure is described in Table 2 of Valdés Hernández et al. (2015a) and single stroke class was created by combining recent and old. All images were re-sliced as to have 1mm in both dimensions of axial slices, with the remaining dimension (slice thickness) left unchanged. White matter probability maps were obtained from T1 image segmentation using Ledig et al. (2015) and cerebro-spinal track probability maps were obtained by co-registering a tract probability map (Hua et al., 2008) to the FLAIR image space. Additionally, in order to have consistent intensity voxel values for model training all MR images were normalized as to have zero mean and standard deviation of one (excluding the background). Values below three standard deviations from the mean clipped in order to guarantee consistent background values across all images.

4. Experiments and results

Data used as input for training the proposed CNN (uResNet) was sampled from whole brain MR volumes as explained in Section 2.3.
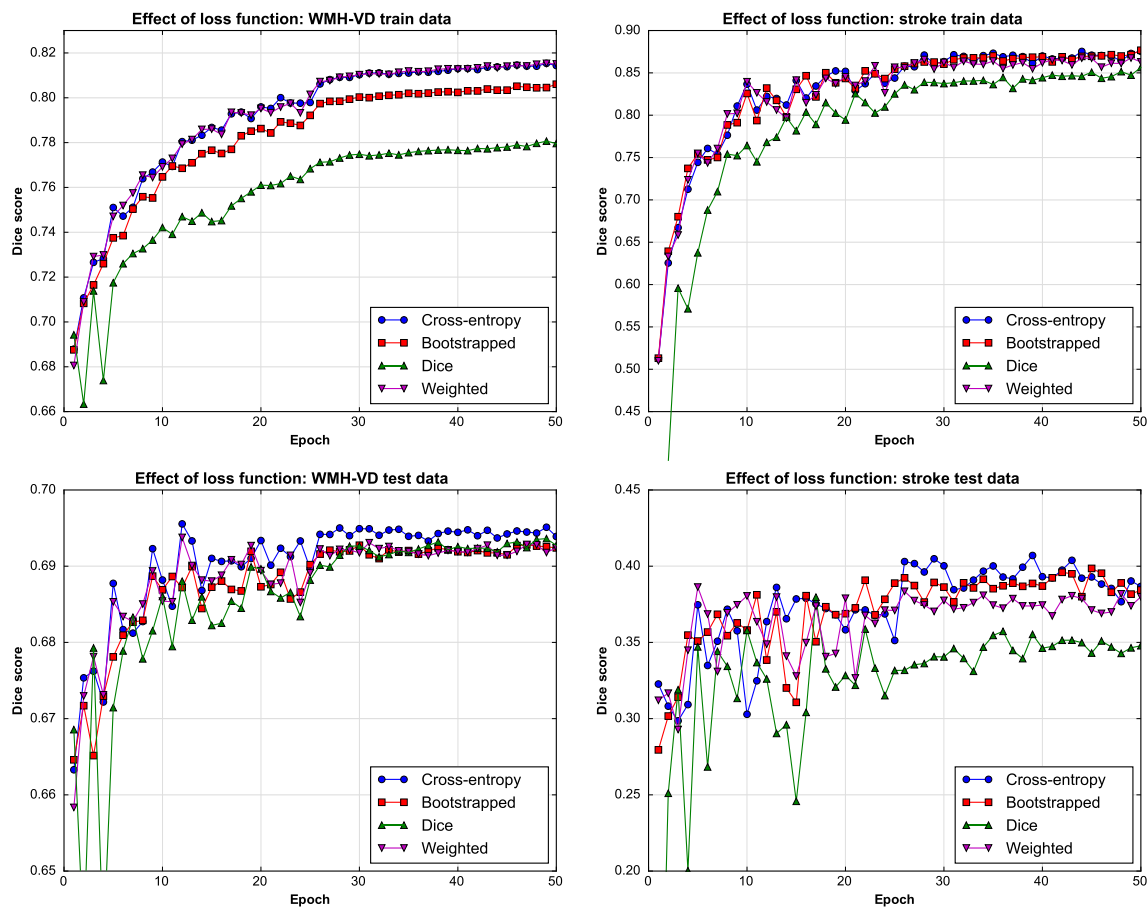
Fig. 3. Effect of different loss functions on uResNet trained using FLAIR images.

Image patches and the corresponding labels of 64×64 voxels were extracted from the volumes at a random subset of 20% of all possible locations that were labeled as WMH and 80% of locations labeled as stroke. All 167 images included in this study contained WMH lesions, of these, 59 also contained stroke lesions. Data was split into two separate sets used for twofold cross-validation, where each fold contained half of the images with WMH only and half with both WMH and stroke, as to represent data distribution in both folds. During each fold of the cross validation experiments one fold is used for training (network parameter learning) and setting all other parameters, while the second (unseen) fold is reserved for testing. That is, optimization of the loss function, input channel selection and stopping criteria are carried out on the training set. Appendix A for a comparison of the proposed uResNet with a version that used residual blocks with two convolutions (uResNet2) and to observe the added value of the center shifting during training patch sampling (uResNet_NoC). Experiments were carried out using the Theano (Al-Rfou et al., 2016) and Lasagne (Dieleman et al., 2015) frameworks with Adam (Kingma and Ba, 2014) optimization (default Lasagne parameters), mini-batch size of 128, learning rate of 0.0005 (with 0.1 decay after 25 epochs) and random weight initialization (default Lasagne settings).

The evaluation criteria used to compare all methods can be split into two, mainly, a comparison to other well established and state-of-the-art methods and a clinical analysis. The comparison to other methods consisted of an evaluation of the labeling overlap of the segmented images using the Dice coefficient, and an analysis of the differences between the automatically calculated volumes and the expert in terms of intra-cranial volume (ICV). Comparison results calculated using the Dice coefficient and volume analyses are reported on a per class basis. Clinical evaluations consisted of a correlation analysis with some clinically relevant variables (mainly Fazekas and MMSE scores), and a

general linear model (GLM) analysis of association with known risk factors.

4.1. Model training

An important factor during CNN training is the definition of the loss function that will guide the learning process (Section 2.2). Here, we experimented with several recently proposed loss functions that were used to train the proposed WMH segmentation CNN using FLAIR images as input. In order to directly compare the effect of different loss functions, Dice score results from evaluating the CNN after different stages of training were calculated, see Fig. 3. Here, the horizontal axis indicates number of training epochs while the vertical axis indicates the Dice score achieved on either the train (top row) or test (bottom row) datasets. In this work, an epoch is defined as transversing the complete set of training of patches once. It must also be noted that Dice results displayed here are calculated on the whole brain MR volumes, not on the extracted patches. Fig. 3 shows the results obtained using classical, bootstrapped and weighted cross-entropy loss functions, as well as using a pseudo Dice similarity score (see Section 2.2). From the top row of Fig. 3 (results on train data) it can be observed that weighted and classical cross-entropy perform best and that there is little difference between them. However, weighted cross-entropy has an additional class weight parameter associated that also need to be set. Hence, for the problem presented in this work and considering the experiments conducted, classical cross-entropy was considered the best choice. It is important to take notice that using the Dice coefficient as both loss function and evaluation metric provides surprisingly poor results during training (top row Fig. 3). Here, we theorize that, for this particular problem, the solution space over which we optimize might be more complex for the Dice metric than the other, and hence finding a
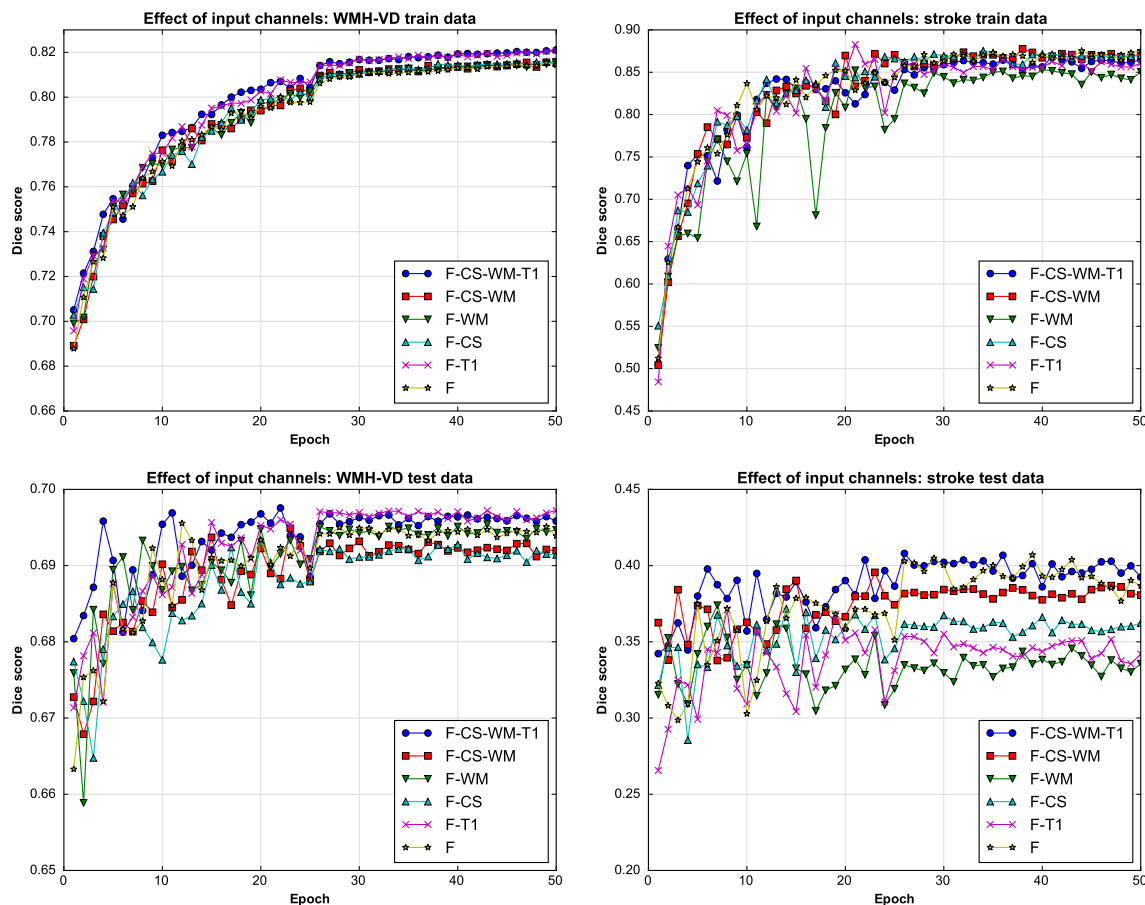
Fig. 4. Different input channel exploration. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.

global optimal solution might prove more cumbersome.

As mentioned before, WMHs are best observed in FLAIR MR images, however it has been suggested that complementary information might be found on T1 MR images. In this work, the contribution from additional multi-modal information to the proposed segmentation framework was explored. Additional input channels to the proposed CNN include T1 MR images, white matter probability maps and a cerebro-spinal tract atlas. Segmentation accuracy is again evaluated using the Dice score. From Fig. 4, it can be seen that training converges after about 30 epochs, that is, traversing the whole set of extracted training patches 30 times. Therefore, test Dice scores and automatic volumes further presented here are obtained evaluating the model at 30 epochs.

Given the different input channels, training data and testing results that take into account both segmentation classes (shown in Fig. 4) indicate that there is very little difference between using four input channels (FLAIR, T1, WM and CS) compared to just using FLAIR images. Hence, all subsequent experiments made use of only FLAIR images as input channels. This is additionally justified by the fact that some of the comparison methods only use FLAIR images. Furthermore the acquisition of additional modalities (T1) or probability map generation (WM and CS) can be costly/time consuming and render the methodology less clinically viable. In Figs. 3 and 4 it can also be observed that training and testing Dice scores for stroke segmentations are much more oscillatory than those from WMH segmentation. This behavior can be explained by the fact that there is simply a lot less data of the stroke class, in fact there are ~14 times more WMH voxels. Therefore, stroke results are more sensitive to variations in the network's parameters as each epoch provides more stochastic gradients associated to this class. Furthermore, the stroke higher training accuracy combined with the lower test accuracy can be attributed to this class imbalance as they potentially point to an over-fitting problem.

4.2. Comparison to state-of-the-art

In the experiments presented in this section the proposed uResNet segmentation CNN was compared to other well established and state-of-the-art algorithms. From the lesion segmentation toolbox (LST) version 2.0.15 (http://www.statistical-_modelling.de/lst.html) the LPA and LGA frameworks were used. LPA was used using only FLAIR images as input while LGA required both FLAIR and T1 images. DeepMedic, a recently published CNN library for segmentation of medical images, was also used in the comparisons presented here with its default settings. Parameters for both LPA and LGA frameworks were set according to a twofold cross-validation using the same data splits as described before for uResNet. LPA has only one parameter, a threshold τ used to binarize the lesion probability maps generated, and the optimal value τ after cross-validation was set to 0.16. The authors recommend setting this value to 0.5, however this produced poor results and hence were excluded from further analysis. LGA has two parameters, κ that was set to 0.12 after cross-validation and a threshold that was set to 0.5. DeepMedic was also validated using the same twofold cross-validation strategy (with FLAIR images as input), where the network is trained in one fold and tested in the other, however, no other meta-parameter (e.g. network's filter sizes, number of feature maps or learning rate) tuning was done. DeepMedic was trained using images re-sliced to isotropic $1mm^3$ voxel size, and patch sampling was internally handled by DeepMedic. The default sampling option was used, which randomly samples 50% of patches from healthy tissue and 50% from pathological tissue (without considering different class weightings).

Dice overlap scores between automatically generated and expertly annotated WMH and stroke lesions are shown in Table 2. Here, it can be observed that the proposed uResNet outperforms the compared methods, with all comparisons between the Dice scores obtained with

Table 2

Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables.

	uResNet	DeepMedic	LPA	LGA	Expert
WMH Dice (std)	**69.5(16.1)**	66.6(16.7)	64.7(19.0)	41.0(22.9)	–
Stroke Dice (std)	**40.0(25.2)**	31.3(29.2)	–	–	–
WMH R^2	**0.951**	0.943	0.855	0.687	–
Stroke R^2	**0.791**	0.688	–	–	–
WMH Trend	0.89x + 0.07	**0.91**x-0.06	0.83x + 0.28	0.51x + 0.16	–
Stroke Trend	**0.58x** + 0.01	0.52x-**0.00**	–	–	–
CC D-Fazekas	**0.770**	0.769	0.746	0.630	0.774
CC PV-Fazekas	0.778	**0.780**	0.777	0.718	0.765
CC Fazekas	**0.824**	**0.824**	0.811	0.734	0.819
CC MMSE	0.364	0.369	**0.443**	0.389	0.372

the proposed and every competing being found to be statistically significant $p < 0.01$ according to Wilcoxon's signed rank test. Statistical significance gives a measure of the likelihood that the difference between two groups could be attributed to change, while effect size (or the "strength of association") quantifies the relative magnitude of the difference between those two groups. Cohen (1988) describes effect size values of 0.2, 0.5 and 0.8 as small, medium and large, respectively. Effect sizes related to the statistical significance tests were calculated with $z/\sqrt{n_1 + n_2}$ as suggested by Pallant (2010), and were 0.45, 0.32 and 0.61 for the comparison of uResNet Dice scores against those from DeepMedic, LPA and LGA, respectively. Fig. 5 shows a correlation analysis between the expertly annotated WMH volumes and those automatically generated. To remove any potential bias associated with head size and thus allow a better comparison, volumes were converted to ICV %. Ideally, automatic algorithms should produce values as similar as possible to the expert, and hence, should lie close to the dotted lines in Fig. 5. The solid lines indicate the general linear trend of the

expert vs. automatic comparison and the coefficient of determination R^2 indicates to what degree automatic values explain the expert ones. From Fig. 5 bottom row we can see that both LPA and LGA perform clearly worse than the CNN approaches (uResNet and DeepMedic). It is also evident that LPA outperforms LGA, where each has a R^2 value of 0.86 and 0.69, respectively. Differences between uResNet and DeepMedic (top row of Fig. 5) are less evident. However, on close inspection of the R^2 metric in Table 2 of uResNet and DeepMedic we can see that uResNet results are slightly better correlated to those generated by the expert. On the other hand, DeepMedic has a slope of 0.91 (offset 0.06) while uResNet has a slope of 0.89 (offset 0.07), suggesting a slightly better agreement.

Fig. 6 shows Bland-Altman plots that further compare expert and automatic WMH volumes. In these plots, the horizontal axis gives the average between expert and automatic volumes for each subject, while the vertical axis shows the difference between these volumes. The reproducibility coefficient (RPC), as calculated here, gives a measure of the variability (or spread) of the differences between automatic and manual volumes and is calculated as 1.96 times the standard deviation σ of those differences (1.96 * σ). In the experiments presented here, smaller values indicate better agreement between automatic and manual volumes. The coefficient of variation (CV) is given by $100*\sigma/\overline{X}$, where $\overline{X}$ refers to the mean volume from both measurements. Dotted lines in the plots of Fig. 6 give the range of the RPC. Bland-Altman plots also provide insight into possible biases of compared methods. LGA displays a statistically significant ($p = 0.85$ to reject zero mean hypothesis) tendency to under-estimate volumes (central solid line). However, all methods tend to under-estimate larger volumes and over-estimate small ones, with the effect more pronounced in LGA.

One of the main objectives of the work presented here is to also differentiate between WMH and stroke lesions. Neither LPA nor LGA are capable of making such a distinction, and therefore are not suitable algorithms for this problem. Fig. 7 (top-row) shows the correlation analysis between automatic (uResNet and DeepMedic) and expert
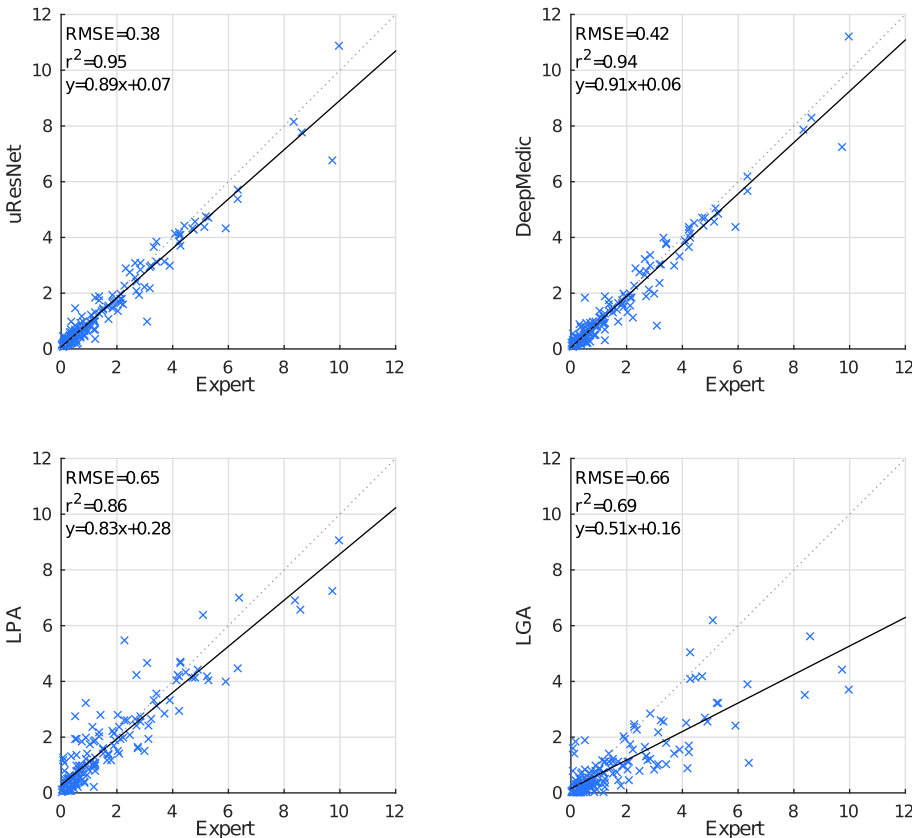

Fig. 5. Automated versus expertly generated WMH volumes (as ICV %). The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted line indicates the ideal trend $f(x) = 1.0x + 0.0$.
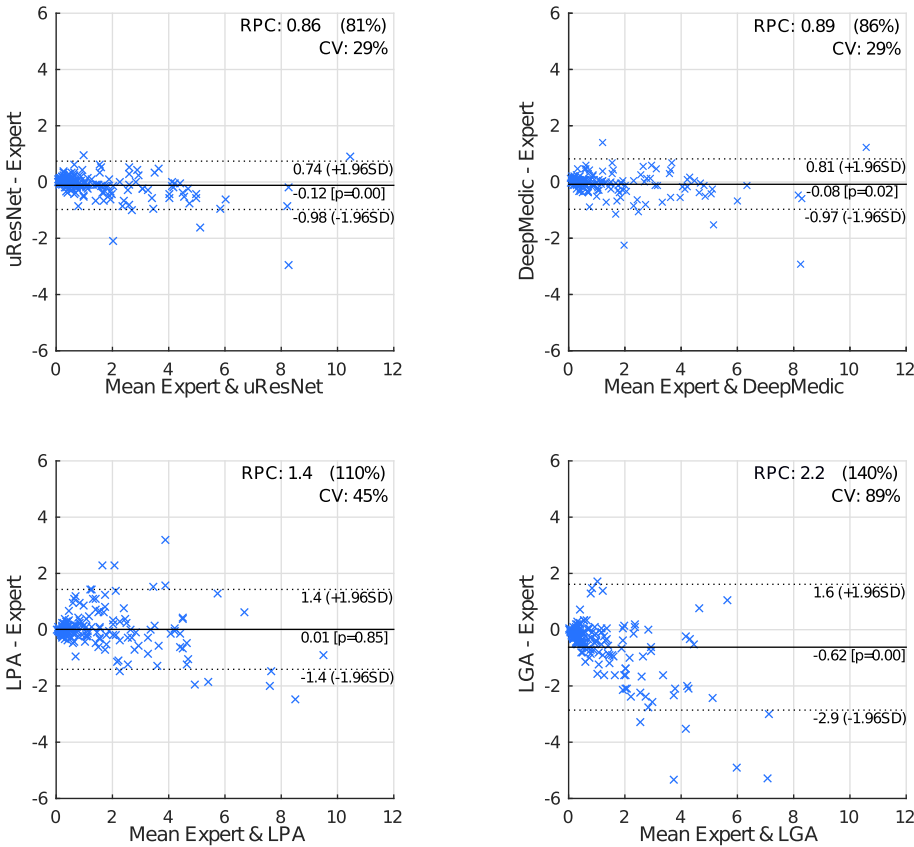
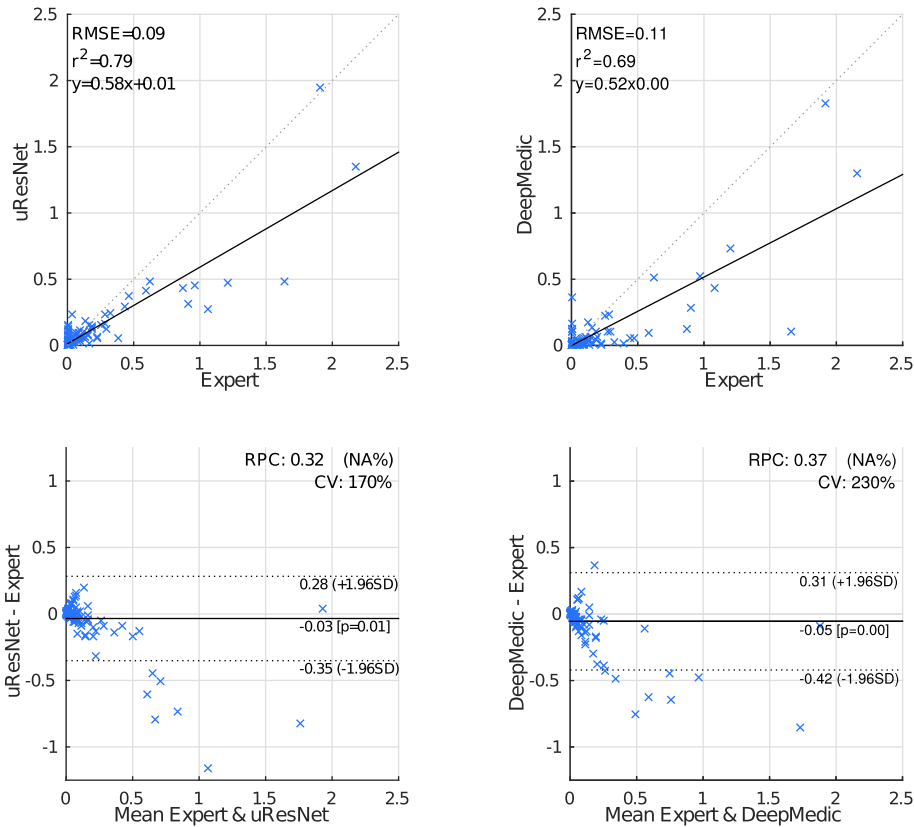**Fig. 6.** Bland-Altman plots comparing expert annotations with all other methods in WMH segmentation.

Fig. 7. Automated versus expertly generated stroke volumes. LPA and LGA are unable to distinguish between WMH and stroke, hence cannot generate these results. The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted indicates the ideal trend $f(x) = 1.0x + 0.0$.
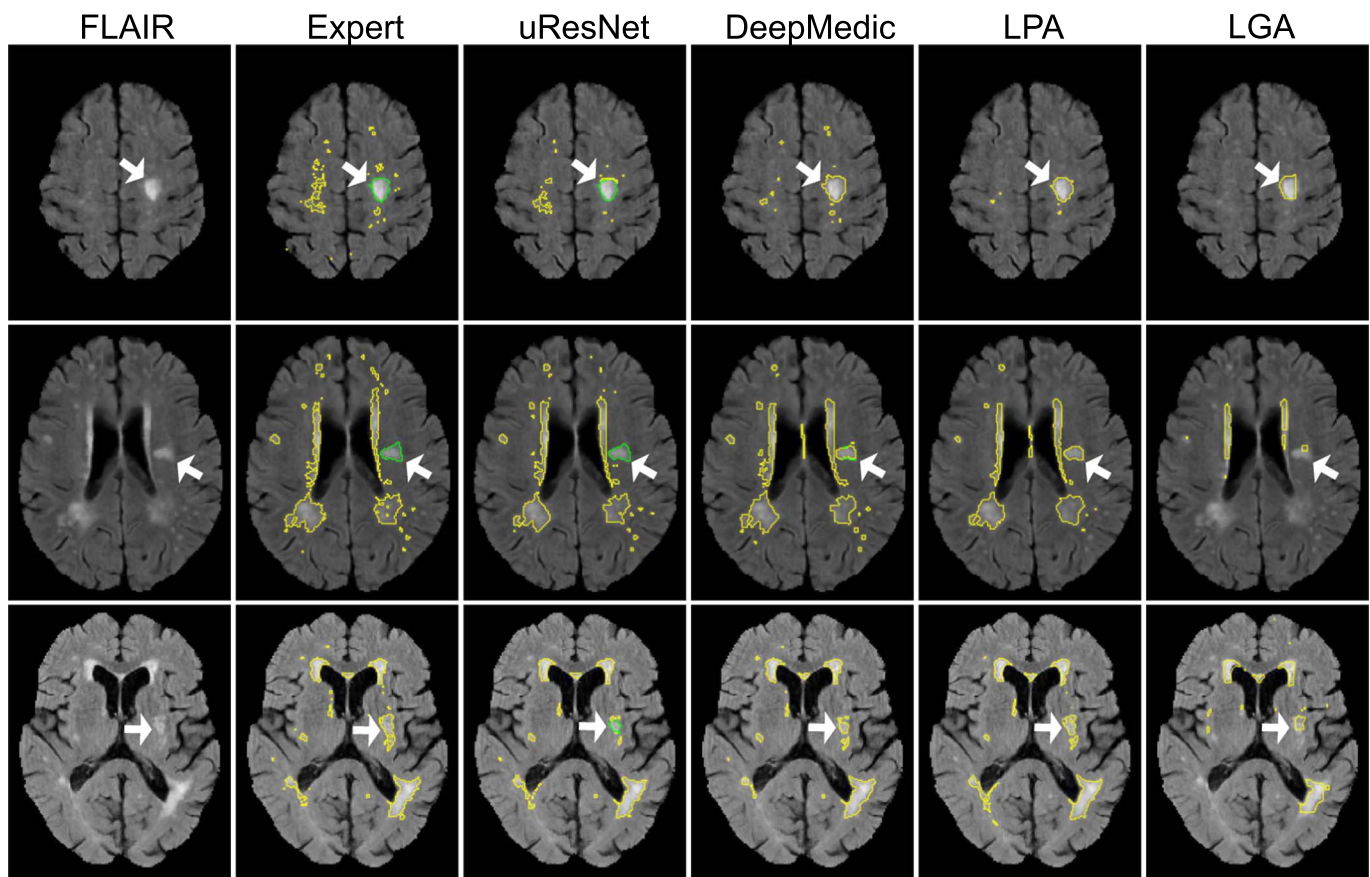
Fig. 8. Visual comparisons of all competing methods. Yellow lines delineate WMH, green lines stroke and white arrows point to interesting result areas. Best seen in color.

stroke volumes (normalized as ICV). It is evident that uResNet out-performs DeepMedic in terms of RMSE, R^2 and linear fit slope. Further to this analysis, Fig. 7 (bottom-row) shows Bland-Altman plots that further confirm these findings, where uResNet obtains a smaller RPC and CV than DeepMedic, with neither method on average displaying a statistically significant tendency to over- or under-estimate volumes (see central solid line on plots). However, it is worth noting that both methods have a tendency to over-estimate small volumes and under-estimate larger ones. A summary of Figs. 5 and 7 is also presented in Table 2, where a difference between both algorithms in terms of Dice scores can be observed. Statistical significance between the comparison of uResNet and DeepMedic Dice scores was found to be $p < 0.05$ according to Wilcoxon's signed rank, with an effect size related to this statistical significance (as suggested by Pallant (2010)) of 0.12. The gap between uResNet and DeepMedic can be considerably closed if additional inputs are provided to DeepMedic (see Appendix B), however this requires an additional MR image acquisition (and co-registration of such image), tissue segmentation and/or co-registration of a cerebro-spinal track atlas. Furthermore, in Appendix C results of DeepMedic experiments that aim to approximate the sampling scheme used by uResNet are discussed.

Fig. 8 shows the segmentation results from three example subjects that illustrate the differences between the methods. Here, it can be observed that uResNet generally does a better job at differentiating between WMH and stroke lesions when compared to DeepMedic (top and middle row). In the bottom row of Fig. 8 an example is illustrated when uResNet wrongly segments some WMH as stroke. Additionally, in the top row, all methods are shown to clearly under-segment the image when compared to the expert is shown. However, inspecting the FLAIR image of this subject (top row, leftmost column) it can be seen that the under-segmented regions would be challenging even for another expert annotator.

4.3. Clinical evaluation

Experiments thus far indicate a better agreement between volumes generated by uResNet and expert annotations, however, the question of the clinical validity of such results remains open. In this regard, Table 2 gives correlation coefficient (CC) results between the volumes and some clinical variables (Fazekas scores and MMSE). Fazekas scores were split into deep white matter (D-Fazekas) and peri-ventricular (PV-Fazekas), with values ranging from 0 to 3. An additional combined Fazekas score, created by adding both D-Fazekas and PV-Fazekas, is also presented. From Table 2 we can observe that in terms of correlation to Fazekas score the proposed uResNet outperforms the other competing methods, additionally noting that CC results for PV-Fazekas and Fazekas are even higher than those obtained from the expert annotations. However, in terms of CC with MMSE it was LPA that performed best.

Using the clinical scores as well as known risk factors available, an analysis of association between WMH volumes and risk factors was carried out. In order to explore such associations a GLM between the results of every algorithm (as well as the expert) and the risk factors was generated. In these models, the risk factors and clinical scores were treated as dependent variables, while the volumes acted as the independent variable. After careful consideration, age, sex, reported diabetes, reported hypertension, reported hyperlipidaemia, reported smoking, total cholesterol, deep atrophy volume and BGPVS score were used in the GLM analysis. Table 3 provides p-values that indicate if a particular risk factor associated with the generated WMH volumes, where the GLMs were corrected for gender differences. Results indicate that only BGPVS is found to be associated with the expertly generated volumes, however deep atrophy volume was also found to be associated with all other methods. Additionally, LPA volumes were also found to be associated with age and diabetes.

In GLM analysis, values that are not well described by the model

Table 3
P-values of linear regression associations between volumes calculated with different methods and risk factors. Bold numbers indicate statistical significance above 0.05.

	uResNet	DeepMedic	LPA	LGA	Expert
Age	0.491	0.533	**< 0.001**	0.723	0.313
Diabetes	0.082	0.072	**0.003**	0.070	0.066
Hyperlipidaemia	0.645	0.547	0.551	0.687	0.728
Hypertension	0.820	0.781	0.504	0.358	0.562
Smoking	0.497	0.560	0.216	0.719	0.767
totalChl	0.235	0.281	0.161	0.328	0.371
BGPVS	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
deepAtrophyVol	**0.015**	**0.019**	**< 0.001**	**< 0.001**	0.117

Table 4
P-values of linear regression associations between volumes calculated with different methods and risk factors after residual outliers were removed. Bold numbers indicate statistical significance above 0.05.

	uResNet	DeepMedic	LPA	LGA	Expert
Age	0.905	0.993	**< 0.001**	0.685	0.407
Diabetes	**0.012**	**0.019**	**< 0.001**	0.177	**0.003**
Hyperlipidaemia	0.346	0.425	0.464	0.550	0.186
Hypertension	0.639	0.502	0.190	0.128	0.350
Smoking	0.069	0.084	0.107	0.673	0.343
totalChl	0.294	0.212	0.222	0.043	0.868
BGPVS	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
deepAtrophyVol	**0.005**	**0.008**	**< 0.001**	**< 0.001**	**0.020**

(outliers) can have a significant impact in subsequent analyses. Outliers in GLM can be identified by examining the probability distribution of the residuals. In order to eliminate any potential bias introduced by outliers, an analysis with outliers removed was performed. Results of this outlier-free association analysis are presented in Table 4. Fig. 9 shows the normal probability plot of residuals for all methods before and after outlier removal. From Table 4 we can observe that once outliers were removed, expert volumes were found to be associated with deep atrophy volume, BGPVS and diabetes. The same associations were found for uResNet, DeepMedic and LPA, with the addition that LPA was again also associated with age. LGA was found to only be associated with BGPVS and deep atrophy volume.

Fazekas scores are highly co-linear with WMH volume (the dependent variable) and therefore were excluded from all previous GLM analysis. Nonetheless, a GLM that included Fazekas scores was also composed as a sanity check that the correct associations would be found. A single Fazekas score was generated by adding the D-Fazekas and PV-Fazekas scores (0–6 score scale). All models found a very strong association ($p \ll 0.001$) between Fazekas and WMH volumes. The effect size for the association of Fazekas score with the expertly generated WMH volumes indicates that a change of one in the Fazekas scale, translates to a change of 0.75 ICV % increase of WMHs (1–0.75). DeepMedic obtained the closest effect size of the association between Fazekas scores and WMH volumes to that of the expert, with a prediction that an increase of one Fazekas point produces 0.70 ICV % increase of WMH (1–0.70). uResNet closely followed with 1–0.69 predictions. LPA and LGA results produced effect sizes of 1–0.6 and 1–0.35, respectively. Of the expert stroke lesion volumes, systolic blood pressure was the only risk factor to be found associated ($p < 0.05$), which incidentally was also associated with the automatically (uResNet and DeepMedic) generated volumes. uResNet values were additionally found to be associated with hypertension. However, it is important to note the small size and heterogeneous nature of the population used in this analysis, which might not prove sufficient to uncover some associations. Due to the small sample analyzed no outlier removal analysis was performed for stroke associations.

5. Discussion

In this work we have proposed a CNN framework, uResNet, for the segmentation of WMHs that is capable of distinguishing between WMHs arising from different pathologies, mainly WMHs of presumed VD origin and those from stroke lesions. Comparison results indicate that the proposed uResNet architecture outperforms other well established and state-of-the-art algorithms.

The architecture used in uResNet follows closely the architecture of U-Net Ronneberger et al. (2015). The main difference being the use of residual elements and a generally lower complexity through the use of summation instead of concatenation in skip connections. Preliminary experiments with both summation and concatenation of features maps found no difference in performance, hence low complexity was favored. However, it is also noted that a more general solution is given by the use of concatenation, as this would allow the network to learn which is the best way of combining the feature maps during training. Of course this additional complexity comes at the expense of a higher risk of overfitting and a higher memory consumption. As mentioned, the use of residual units provides advantages during training, mainly improved convergence rates in our experiments. Recently, He et al. (2016b) proposed a new pre-activated residual unit, which optimizes the architecture of each unit making training easier and improving generalization. Future work will involve updating the architecture to include such residual elements and evaluating their merits in the context of WMH segmentation.

Large class imbalance in medical image segmentation is generally an issue that must be considered. Loss functions that take into account the class imbalance have the drawback that they have the additional class weighting parameter to tune. An additional complication resulting from a large class imbalance is that a lot of computational effort might be spent optimizing to perform well in large and relatively easy to classify/segment sections of an image. Bootstrapped cross-entropy attempts to focus the learning process on hard to classify parts of an image by dropping out loss function contribution from voxels that have already been classified to a good degree of certainty. However, this technique also requires the setting of an additional parameter, the threshold to consider a classification as already good, and moreover, evaluation results indicated a performance similar to classical cross-entropy.

A very important factor of the proposed CNN framework is the training data sampling strategy described in Section 2.3. CNN training for medical imaging using patches is a somewhat standard technique that helps reduce the very large class imbalance that usually affects medical image segmentation. However, careful consideration must be given in the sampling strategy adopted for a certain architecture. As mentioned, class imbalance and lesion location within samples need to be considered. The use of the proposed sampling strategy described in Section 2.3 had a profound effect on the proposed uResNet, with WMH and stroke Dice scores increasing from ∼67 to ∼70 and from ∼29 to ∼40, respectively, due to this alone. Another important factor is the frequency each class is sampled. In this work we sampled at 20% of the locations labeled as WMH while at 80% of the locations labeled as stroke, again to try to balance classes. It is important to note that the default sampling settings of DeepMedic were used as in Kamnitsas et al. (2017). In this default sampling strategy, DeepMedic samples equally from healthy and diseased tissues (that is without, considering frequency of different diseased classes) and furthermore does not include the central voxel offset sampling strategy used here. We believe both these factors had a significant impact in the differences between these methods, specially in the stroke lesion class. Training data was augmented by applying random flips to training patches, however we did not find this had a clear effect on results.

An important aspect to note is that WMH segmentation is notoriously challenging: For example, Bartko (1991) and Anbeek et al. (2004b) consider similarity scores of 70 to be excellent, while Landis
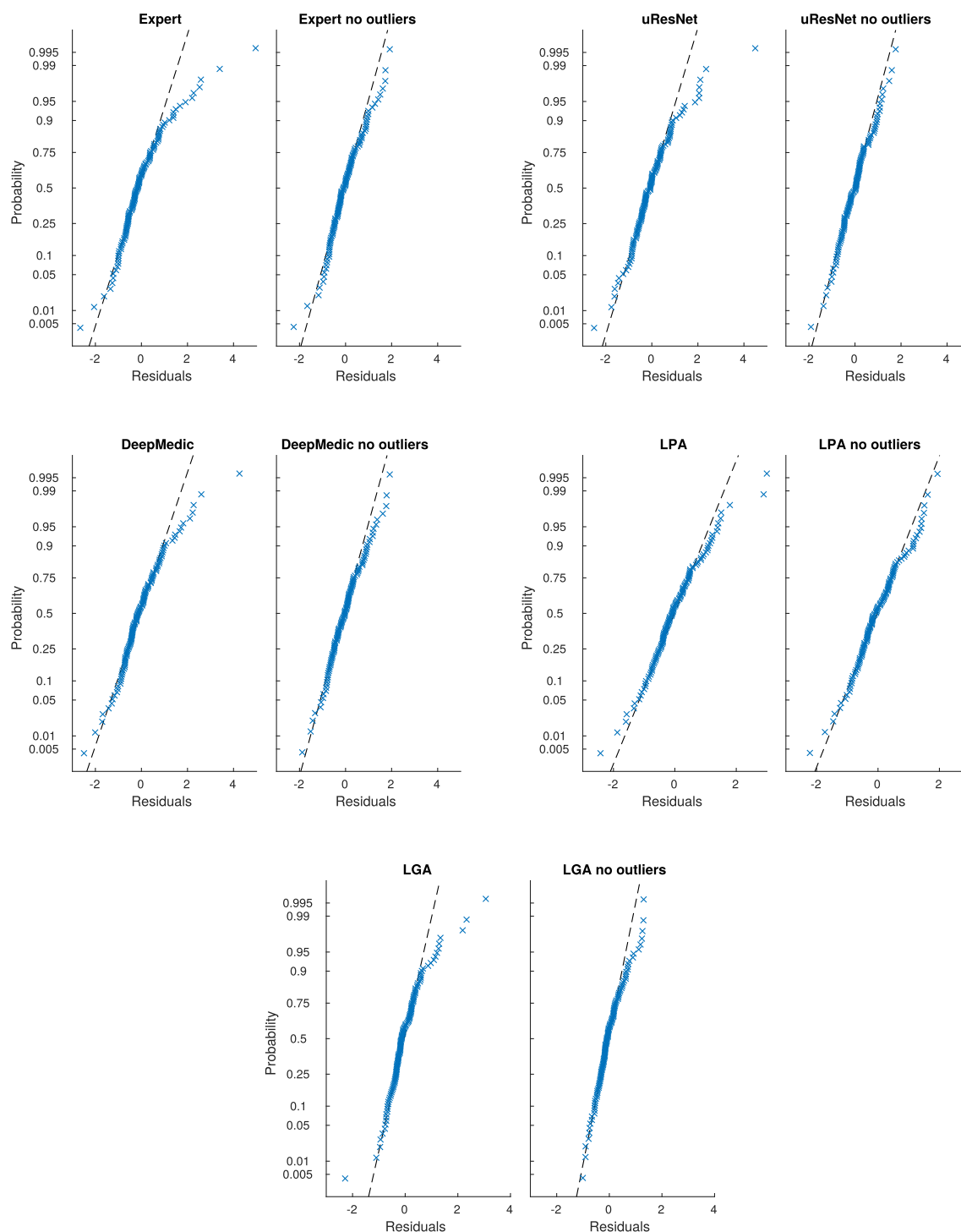
Fig. 9. General linear model normal probability plots of residuals for all methods, with and without outliers.

and Koch (1977) consider scores of 40, 60 and 80 to be moderate, substantial and near perfect, respectively. With this in mind, we can consider average Dice scores for WMHs generated by the proposed uResNet, as well as those from DeepMedic and LPA to all be substantial, with LGA generating only moderate results. It is important to note that LGA is at heart an unsupervised method and that data was only used to tune its κ parameter. Only uResNet and DeepMedic are capable of distinguishing between different types of lesion, and in this regard only uResNet produced an average stroke Dice score that could be considered moderate.

Appendix A. Variations of uResNet

In this section we present results comparing the proposed architecture and sampling scheme, with two additional version: One where the residual block takes the more traditional form of two convolutional elements (called uResNet2) and another where the proposed center shifting sampling scheme is replaced with a standard centered patch sampling scheme (called uResNet_NoC, for not off-centered). Table A.5 summarizes these results.

Using single convolution residual blocks was noted (He et al., 2016a) to be equivalent to a linear projection. After experimented with residuals blocks of one and two convolutions, we observed no statistical difference ($p > 0.05$) between them. However, learning the residual of these linear projections might still be simpler, thus leading to an observed faster convergence. This observations need to be interpreted with care. We believe that the Dice overlap scores that our method achieves are close to expected intra-rater variability, hence the lack of observed difference in performance between one and two convolutions in residual blocks, might come down to limitations of the data itself.

Training with patches that always contain a diseased label in the center would bias towards labeling this region of a patch as diseased during inference. Patch center shifting alleviates this problem due to the distribution of probability to observe a lesion across the whole field-of-view. For example, if we would estimate the probability of observing a lesion in any particular location of a training patch, there would be 100% probability to observe a lesion at its center (Fig. A.10 (a)), as we explicitly sampled in this manner. Allowing patches to be shifted spreads this probability to all locations and not any single location has a preferential likelihood of being a lesion (Fig. A.10 (b)). In a fully convolutional neural network predictions can be made over a large area (as the network proposed here), taking into account context information from large areas of an image (the field-of-view or receptive field). However, training is driven by pixel-wise prediction errors, hence labeling occurs on a per-pixel basis. The likelihood of observing a lesion at any particular location is in fact very low (see Fig. A.10 (b)) and more or less uniform. It is this uniformity that removes the bias towards any particular location. Results comparing a uResNet with out center shifting sampling are shown in Table A.5.

Table A.5
Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables. No statistical significance between uResNet and uResNet2 was observed ($p > 0.05$), while there was a statistically significant difference ($p < 0.001$) between patch off-center sampling (uResNet) and regular no off-center sampling (uResNet_NoC).

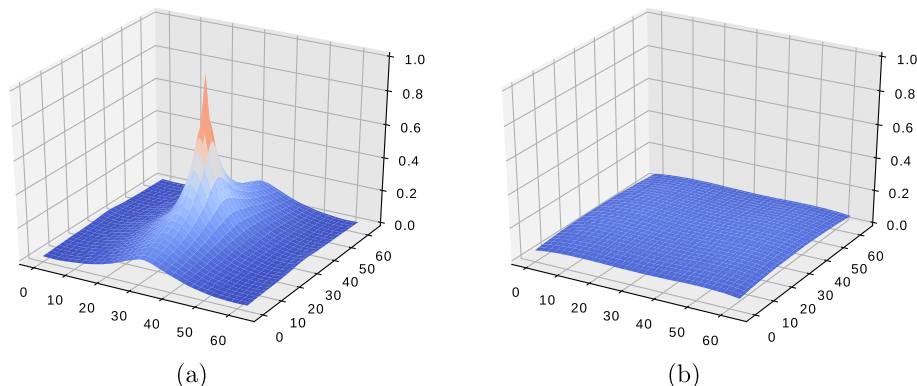|	uResNet	uResNet2	uResNet_NoC	Expert
WMH Dice (std)	69.5(16.1)	**69.6(16.1)**	66.9(18.1)	–
Stroke Dice (std)	40.0(25.2)	**40.2(27.7)**	28.9**(22.3)**	–
WMH R^2	**0.951**	**0.951**	0.948	–
Stroke R^2	**0.791**	0.761	0.710	–
WMH Trend	**0.89x + 0.07**	**0.89x**-0.08	**0.89x** + 0.15	–
Stroke Trend	**0.58x + 0.01**	0.55x-**0.01**	0.52x + 0.07	–
CC D-Fazekas	0.770	**0.776**	0.771	0.774
CC PV-Fazekas	0.778	**0.783**	0.777	0.765
CC Fazekas	0.824	**0.831**	0.823	0.819
CC MMSE	0.364	0.373	0.366	0.372

Fig. A.10. Lesion likelihood on training patches (a) without shifting and (b) after shifting.

Appendix B. Dice results for different inputs

Both CNN approaches, uResNet and DeepMedic, can easily be trained using one or several inputs. Table B.6 provides Dice overlap results of using different input channels in both CNN approaches. As it can be appreciated DeepMedic can narrow the Dice overlap gap with uResNet if several inputs are provided. However, as discussed before, obtaining and generating these extra inputs limit clinical applicability and also add additional computational costs to the whole segmentation framework.

Table B.6

Mean Dice scores of WMH and stroke, for different inputs with uResNet and DeepMedic. Difference in Dice score between the two methods is given in italics. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.

Input	WMH			Stroke		
channels	uResNet	DeepMedic	Diff	uResNet	DeepMedic	Diff
F	**69.5**	66.3	*3.2*	**40.0**	31.1	*8.9*
F-T1	**69.7**	67.6	*2.1*	**35.5**	34.3	*1.2*
F-CS	**69.1**	66.6	*2.5*	**36.7**	35.1	*1.6*
F-WM	**69.4**	68.2	*1.2*	33.0	**35.9**	*− 2.9*
F-CS-WM	**69.3**	68.0	*1.3*	**38.4**	37.8	*0.6*
F-T1-CS-WM	**69.6**	68.4	*1.2*	**40.2**	36.0	*4.2*

Appendix C. Additional DeepMedic experiments

DeepMedic experiments that aim to approximate the sampling scheme used by uResNet were carried out, where several sampling weights were tested for DeepMedic. A direct comparison of per-class patch sampling is not straightforward between the proposed method and DeepMedic, and furthermore it can be misleading. For instance, in the work proposed here a sampling rate of 80–20% of WHM-stroke patches is used, each patch has a size of 64 by 64 voxels and uResNet makes a prediction of a 64 by 64 patch of the label space during training (it is fully convolutional and uses padded convolutions throughout). This means that each patch used in uResNet has a label map that due to its size inevitably contains a large amount of healthy tissue. Therefore we do not sample specifically from healthy regions. On the other hand, DeepMedic trains with segments that have a label space of 9 by 9 by 9 voxels, therefore it is far less likely that healthy tissue is included in non-healthy samples and thus healthy segments need to be sampled. Nonetheless, different per-class sampling rates, as well as other hyper-parameter settings with DeepMedic were explored.

Some of DeepMedic's default hyper-parameter values are: learning rate of 1e-3, RmsProp optimizer, sampling form of foreground/background (diseased/healthy tissue) and sampling rate of [0.5, 0.5] (healthy and diseased tissue). The different sampling rates tested with DeepMedic in our experiments to approximate uResNet setup were [0.5, 0.1, 0.4], [0.5, 0.25, 0.25], [0.33, 0.13, 0.53] and [0.33, 0.33, 0.3], for healthy, WMH and stroke tissue, respectively. Additionally, learning rate values explored were in the range of 1.9e-2 to 1e-4, with RMSprop, Adam or SGD as optimizer. Changing the sampling rates from the default generally produced unstable results, with either failing to converge or producing poorer overlap values than with the default settings. In total, 14 different additional DeepMedic train/test runs were performed, out of which only two converged, both using a sampling rate of [0.33, 0.33, 0.33]. Dice overlap results by these experiments were of 60.7 and 29.9, for WMH and stroke, respectively, in one instance and 59.4 and 29.5 in the other. These unstable results might be due to the tuning of additional meta parameters, such as the optimizer, learning rate or regularization. Therefore, presented DeepMedic results were obtained with default hyper-parameters, which were the best results obtained in our experiments.

References

SPIRIT, 1997. A randomized trial of anticoagulants versus aspirin after cerebral ischemia of presumed arterial origin. The Stroke Prevention in Reversible Ischemia Trial (SPIRIT) Study Group. Ann. Neurol. 42 (6), 857–865.

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Bleecker Snyder, J., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson, A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., Cooijmans, T., Côté, M.-A., Côté, M., Courville, A., Dauphin, Y.N., Delalleau, O., Demouth, J., Desjardins, G., Dieleman, S., Dinh, L., Ducoffe, M., Dumoulin, V., Ebrahimi Kahou, S., Erhan, D., Fan, Z., Firat, O., Germain, M., Glorot, X., Goodfellow, I., Graham, M., Gulcehre, C., Hamel, P., Harlouchet, I., Heng, J.-P., Hidasi, B., Honari, S., Jain, A., Jean, S., Jia, K., Korobov, M., Kulkarni, V., Lamb, A., Lamblin, P., Larsen, E., Laurent, C., Lee, S., Lefrancois, S., Lemieux, S., Léonard, N., Lin, Z., Livezey, J.A., Lorenz, C., Lowin, J., Ma, Q., Manzagol, P.-A., Mastropietro, O., McGibbon, R.T., Memisevic, R., van Merriënboer, B., Michalski, V., Mirza, M., Orlandi, A., Pal, C., Pascanu, R., Pezeshki, M., Raffel, C., Renshaw, D., Rocklin, M., Romero, A., Roth, M., Sadowski, P., Salvatier, J., Savard, F., Schlüter, J., Schulman, J., Schwartz, G., Serban, I.V., Serdyuk, D., Shabanian, S., Simon, E., Spieckermann, S., Subramanyam, S.R., Sygnowski, J., Tanguay, J., van Tulder, G., Turian, J., Urban, S., Vincent, P., Visin, F., de Vries, H., Warde-Farley, D., Webb, D.J., Willson, M., Xu, K., Xue, L., Yao, L., Zhang, S., Zhang, Y., 2016. Theano: A {Python} framework for fast computation of mathematical expressions. arXiv e-prints, abs/1605.02688, may.

Anbeek, P., Vincken, K.L., Van Osch, M.J., Bisschops, R.H., Van Der Grond, J., 2004a. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 21 (3), 1037–1044.

Anbeek, P., Vincken, K.L., van Osch, M.J., Bisschops, R.H., van der Grond, J., 2004b. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 21 (3), 1037–1044.

Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17 (3), 483–489.

Bendfeldt, K., Blumhagen, J.O., Egger, H., Loetscher, P., Denier, N., Kuster, P., Traud, S., Mueller-Lenke, N., Naegelin, Y., Gass, A., Hirsch, J., Kappos, L., Nichols, T.E., Radue, E.-W., Borgwardt, S.J., 2010. Spatiotemporal distribution pattern of white matter lesion volumes and their association with regional grey matter volume reductions in relapsing-remitting multiple sclerosis. Hum. Brain Mapp. 31 (10), 1542–1555 (oct).

Birenbaum, A., Greenspan, H., 2016. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 58–67.

Bowles, C., Qin, C., Ledig, C., Guerrero, R., Hammers, A., Sakka, E., Dickie, D.A., Tijms, B., Lemstra, A.W., Flier, W.V.D., 2016. Pseudo-healthy image synthesis for white matter lesion segmentation. In: International Workshop on Simulation and Synthesis in Medical Imaging. vol. 9968. Springer International Publishing, pp. 87–96.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 35 (5), 1229–1239.

Brosch, T., Yoo, Y., Tang, L.Y., Li, D.K., Traboulsee, A., Tam, R., 2015. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 3–11.

Brott, T., Marler, J.R., Olinger, C.P., Adams, H.P., Tomsick, T., Barsan, W.G., Biller, J., Eberle, R., Hertzberg, V., Walker, M., 1989. Measurements of acute cerebral infarction: lesion size by computed tomography. Stroke 20 (7), 871–875.

Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., Cherubini, A., 2015. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. Neuroinformatics 13 (3), 261–276.

Cardoso, M.J., Sudre, C.H., Modat, M., Ourselin, S., 2015. Template-based multimodal joint generative model of brain data. In: Information Processing in Medical Imaging. vol. 9123. Springer International Publishing, pp. 17–29.

Chard, D.T., Griffin, C.M., Parker, G.J.M., Kapoor, R., Thompson, A.J., Miller, D.H., 2002. Brain atrophy in clinically early relapsing-remitting multiple sclerosis. Brain J. Neurol. 125 (2), 327–337.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Lawrence Erlbaum Associates, pp. 40.

Dalca, A.V., Sridharan, R., Cloonan, L., Fitzpatrick, K.M., Kanakis, A., Furie, K.L., Rosand, J., Wu, O., Sabuncu, M., Rost, N.S., et al., 2014. Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. In: Medical image computing and computer-assisted intervention: MICCAI. International Conference on Medical Image Computing and Computer-Assisted Intervention. vol. 17. pp. 773.

Damangir, S., Manzouri, A., Oppedal, K., Carlsson, S., Firbank, M.J., Sonnesyn, H.,

Tysnes, O.-B., O'Brien, J.T., Beyer, M.K., Westman, E., Aarsland, D., Wahlund, L.-O., Spulber, G., 2012. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. J. Neurol. Sci. 322 (1), 211–216.

Derntl, A., Plant, C., Gruber, P., Wegener, S., Bauer, J.S., Menze, B.H., 2015. Stroke lesion segmentation using a probabilistic atlas of cerebral vascular territories. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 21–32.

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., de Almeida, D.M., McFee, B., Weideman, H., Takács, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K., Liu, C., French, G., Degrave, J., 2015. Lasagne: First release. (Aug).

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38 (2), 295–307.

Dugas-Phocion, G., Ballester, M.A.G., Malandain, G., Lebrun, C., Ayache, N., 2004. Improved em-based tissue segmentation and partial volume effect quantification in multi-sequence brain mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 26–33.

Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., Zimmerman, R.A., 1987. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. AJR Am. J. Roentgenol. 149 (2), 351–356.

Forbes, F., Doyle, S., Garcia-Lorenzo, D., Barillot, C., Dojat, M., 2010. Adaptive weighted fusion of multiple mr sequences for brain lesion segmentation. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, pp. 69–72.

Freifeld, O., Greenspan, H., Goldberger, J., 2009. Multiple sclerosis lesion detection using constrained GMM and curve evolution. J. Biomed. Imaging 2009, 14.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17 (1), 1–18.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage 57 (2), 378–390.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., Sanchez, C., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., Platel, B., 2016. Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. ArXiv e-prints, Oct.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uder, I.W.M., de Leeuw, F.E., Marchiori, E., van Ginneken, B., Platel, B., 2016, apr, apr. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In: International Symposium on Biomedical Imaging. IEEE, pp. 1414–1417.

Gibson, E., Gao, F., Black, S.E., Lobaugh, N.J., 2010. Automatic segmentation of white matter hyperintensities in the elderly using flair images at 3t. J. Magn. Reson. Imaging 31 (6), 1311–1322.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., et al., 2016. Bianca (brain intensity abnormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. Neuroimage 141, 191–205.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (June).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity Mappings in Deep Residual Networks. (mar).

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D.S., Calabresi, P.A., Pekar, J.J., van Zijl, P.C., Mori, S., 2008. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. NeuroImage 39 (1), 336–347.

Ithapu, V., Singh, V., Lindner, C., Austin, B.P., Hinrichs, C., Carlsson, C.M., Bendlin, B.B., Johnson, S.C., 2014. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. Hum. Brain Mapp. 35 (8), 219–4235 (feb).

Itti, L., Chang, L., Ernst, T., 2001. Segmentation of progressive multifocal leukoencephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging. J. Neuroimaging 11 (4), 412–417 (oct).

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B., 2015. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In: ISLES-MICCAI.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Kawata, Y., Arimura, H., Yamashita, Y., Magome, T., Ohki, M., Toyofuku, F., Higashida, Y., Tsuchiya, K., 2010. Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging. Comput. Med. Imaging Graph. 34 (5), 370–376.

Khayati, R., Vafadust, M., Towhidkhah, F., Nabavi, M., 2008. Fully automatic segmentation of multiple sclerosis lesions in brain MR flair images using adaptive mixtures method and Markov random field model. Comput. Biol. Med. 38 (3), 379–390.

Kikinis, R., Guttmann, C.R., Metcalf, D., Wells, W.M., Ettinger, G.J., Weiner, H.L., Jolesz, F.A., 1999. Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects. J. Magn. Reson. Imaging 9 (4), 519–530.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. (dec).

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., Bryan, R.N., Davatzikos, C., 2008. Computer-assisted segmentation of white matter lesions in 3d

MR images using support vector machine. Acad. Radiol. 15 (3), 300–313.

Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F.J., Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D., 2015. Robust whole-brain segmentation: application to traumatic brain injury. Med. Image Anal. 21 (1), 40–58.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Löuvbld, K.-O., Baird, A.E., Schlaug, G., Benfield, A., Siewert, B., Voetsch, B., Connor, A., Burzynski, C., Edelman, R.R., Warach, S., 1997. Ischemic lesion volumes in acute stroke by diffusion-weighted magnetic resonance imaging correlate with clinical outcome. Ann. Neurol. 42 (2), 164–170 (aug).

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Med. Image Anal. 35, 250–269.

Maier, O., Schröder, C., Forkert, N.D., Martinetz, T., Handels, H., 2015a. Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS One 10 (12), e0145118.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U., Handels, H., 2014. Ischemic stroke lesion segmentation in multi-spectral mr images with support vector machine classifiers. In: SPIE Medical Imaging, 903504–903504.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U.M., Münte, T.F., Handels, H., 2015b. Extra tree forests for sub-acute ischemic stroke lesion segmentation in mr sequences. J. Neurosci. Methods 240, 89–100.

Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., Tzourio-Mazoyer, N., Alpérovitch, A., Tzourio, C., Mazoyer, B., 2008. An automated procedure for the assessment of white matter hyperintensities by multispectral (t1, t2, pd) MRI and an evaluation of its between-centre reproducibility based on two large community databases. Neuroradiology 50 (1), 31–42.

McKinley, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., Reyes, M., 2016. Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis. In: MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure, pp. 37.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.

Pallant, J., 2010. SPSS Survival Manual: A Step By Step Guide to Data Analysis Using SPSS, 4th edition. .

Prakash, K.B., Gupta, V., Bilello, M., Beauchamp, N.J., Nowinski, W.L., 2006. Identification, segmentation, and image property study of acute infarcts in diffusion-weighted images by using a probabilistic neural network and adaptive Gaussian mixture model. Acad. Radiol. 13 (12), 1474–1484.

Qin, C., Guerrero, R., Bowles, C., Ledig, C., Scheltens, P., Barkhof, F., Rhodius-Meester, H., Tijms, B., Lemstra, A.W., van der Flier, W.M., Glocker, B., Rueckert, D., 2016. A semi-supervised large margin algorithm for white matter hyperintensity segmentation. In: Machine Learning in Medical Imaging. vol. 9352. Springer International Publishing, pp. 104–112.

Rekik, I., Allassonnière, S., Carpenter, T.K., Wardlaw, J.M., 2012. Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal. NeuroImage Clin. 1 (1), 164–178.

Ronneberger, O., P.Fischer,Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 9351. Springer, pp. 234–241.

Schmidt, P., 2017. Bayesian Inference for Structured Additive Regression Models for Large-scale Problems with Applications to Medical Imaging. Ph.D. thesis. Ludwig-Maximilians-Universität München.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. NeuroImage 59 (4), 3774–3783.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010a. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010b. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, pp. 1–14 (sep).

Valdés Hernández, M. d. C., Armitage, P.A., Thrippleton, M.J., Chappell, F., Sandeman, E., Muñoz Maniega, S., Shuler, K., Wardlaw, J.M., 2015a. Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. Brain Behav. 5 (12), 1–8.

Valdés Hernández, M. d. C., Maconick, L.C., Muñoz Maniega, S., Wang, X., Wiseman, S., Armitage, P.A., Doubal, F.N., Makin, S., Sudlow, C.L.M., Dennis, M.S., Deary, I.J., Bastin, M., Wardlaw, J.M., 2015b. A comparison of location of acute symptomatic vs. 'silent' small vessel lesions. Int. J. Stroke 10 (7), 1044–1050.

Valverde, S., Cabezas, M., Roura, E., González-Villa, S., Salvi, J., Oliver, A., Lladó, X., 2016. Multiple Sclerosis Lesion Detection and Segmentation using a Convolutional Neural Network of 3d Patches. MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure. pp. 75.

van Gijn, J., 1998. Leukoaraiosis and vascular dementia. Neurology 51 (3 Suppl 3), S3–S8.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20 (8), 677–688.

Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: Medical Image Computing and Computer-Assisted Intervention. vol. 9349. Springer International Publishing, pp. 677–684.

Wang, X., Valdés Hernández, M.C., Doubal, F., Chappell, F.M., Wardlaw, J.M., 2012. How much do focal infarcts distort white matter lesions and global cerebral atrophy measures?. Cerebrovasc. Dis. (Basel, Switz.) 34 (5-6), 336–342.

Wardlaw, J.M., Pantoni, L., 2014. Sporadic Small Vessel Disease: Pathogenic Aspects. Cambridge University Press, Cambridge, pp. 52–63.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., Black, S.E., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., de Leeuw, F.-E., Doubal, F., Duering, M., Fox, N.C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., van Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B.C.M., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P.B., Dichgans, M., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12 (8), 822–838.

Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 8149. pp. 735–742.

Wu, Z., Shen, C., Van Den Hengel, A., High-performance Semantic Segmentation Using Very Deep Fully Convolutional Networks, ArXiv e-prints.

Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. vol. 8149. Springer Berlin Heidelberg, pp. 606–613.